Received: 28 March 2023 | Revised: 14 August 2023 | Accepted: 1 September 2023

DOI: 10.1002/psp4.13049

ARTICLE

Case–control matching-guided exposure-efficacy relationship for avelumab in patients with urothelial carcinoma

Pooneh Soltantabar[1] | **Ali Alhadab**[2] | **Jennifer Hibma**[3] | **Satrajit Roychoudhury**[1] | **Diane D. Wang**[1] | **Carlo Bello**[1] | **Mohamed Elmeliegy**[1]

[1]Oncology Research and Development, Clinical Pharmacology, Pfizer Inc, San Diego, California, USA

[2]Global Development, Janssen R & D, San Diego, California, USA

[3]Clinical Pharmacology and Bioanalytics, Pfizer Inc, San Diego, California, USA

Correspondence
Mohamed Elmeliegy, Oncology Research and Development, Clinical Pharmacology, Pfizer Inc, 10555 Science Center Dr., San Diego, CA 92121, USA.
Email: mohamed.elmeliegy@pfizer.com

Abstract

Exposure-response (E-R) analyses are an integral component of understanding the benefit/risk profile of novel oncology therapeutics. These analyses are typically conducted using data from the treatment arm to characterize the relationship between drug exposure (low vs. high) and efficacy or safety outcomes. For example, outcomes of patients with lower exposure in the treatment arm (e.g., Q1) might be compared to outcomes of those with higher drug exposure (Q2, Q3, and Q4). Outcomes from the lowest exposure quartile may be also compared to the control arm to evaluate whether the Q1 subgroup derived clinical benefit. However, the sample size and the distribution of patient baseline characteristics and disease risk factors are not balanced in such a comparison (Q1 vs. control), which may bias the analysis and causal interpretation of clinical benefit in the Q1 subgroup. Herein, we report the use of case–control matching to account for this bias and better understand the E-R relationship for avelumab in urothelial carcinoma, a PD-L1 inhibitor approved for the treatment of several cancers. Data from JAVELIN-100 was utilized which is a phase III study of avelumab in first-line maintenance treatment in patients with urothelial carcinoma; this clinical study demonstrated superiority of avelumab versus best-supportive care leading to approval in the United States, Europe, and other countries. A post hoc case-control matching method was implemented to compare the efficacy outcome between Q1 avelumab subgroup and matched patients extracted from the control arm with similar baseline characteristics, which showed a clinically relevant difference in overall survival in favor of the Q1 avelumab subgroup. This analysis demonstrates the importance of accounting for imbalance in important baseline covariates when comparing efficacy outcomes between subgroups within the treatment arm versus the control arm.

Study Highlights

WHAT IS THE CURRENT KNOWLEDGE ON THE TOPIC?

Exposure-response (E-R) analysis for oncology drugs could be confounded by the baseline disease risk factors. When comparing the clinical benefit in patients with low exposure versus the control subjects, these disease risk factors may bias the causal relationship between exposure and clinical benefit.

WHAT QUESTION DID THIS STUDY ADDRESS?

Are the patients with urothelial carcinoma in the low avelumab exposure subgroup benefiting from the treatment compared to the control subjects?

WHAT DOES THIS STUDY ADD TO OUR KNOWLEDGE?

Case-control matching can be used in the context of E-R analysis to balance the distribution of baseline risk factors between groups. Using this approach, the bias introduced by disease risk factors can be addressed and the causal relationship between exposure and clinical benefit can be assessed.

HOW MIGHT THIS CHANGE DRUG DISCOVERY, DEVELOPMENT, AND/OR THERAPEUTICS?

Case-control matching can be implemented to account for the imbalance in important baseline covariates, enabling a direct assessment of the causal effect of the exposure on outcome.

INTRODUCTION

Development of new treatments for oncology indications have traditionally been expedited given the unmet need. However, this often hampers robust dose exploration efforts as the exploration of the relationship between dose/exposure and clinical outcomes (efficacy or safety) are limited to small cohorts within the dose escalation stages with heterogenous patient populations and cancer types. The US Food and Drug Administration's (FDA's) Project OPTIMUS[1] advocates for re-assessing dose selection strategies for oncology compounds and recommends randomized dose-finding trials as a preferred approach for dose optimization given that the maximum tolerated dose (MTD) might not be adequate for targeted therapies.[2,3] Still, a single dosing regimen, the MTD or the highest tested dose, has been traditionally evaluated in pivotal studies.[4,5]

Exposure-response (E-R) characterization is at the heart of supporting dose selection.[6–8] The utility of E-R analysis for pivotal oncology studies is limited as they mainly utilize a single dosing regimen. The exposure range observed in these studies is narrow with most of the exposure variability attributed to between subject pharmacokinetic (PK) variability rather than to different doses. Therefore, these analyses have limitations to provide additional insight whether the selected regimen is optimal to maximize the desired efficacy and minimize unwanted toxicities. However, characterizing the E-R relationship for such studies is still necessary for understanding the relationship between exposure and safety or efficacy end points in the exposure range evaluated.

The characterization of the relationship between drug exposure and other factors (e.g., patient's baseline characteristics, disease status, etc.) with clinical outcomes is often conducted within the treatment arm only (i.e., the control arm is excluded from the E-R analysis given the lack of exposure from the investigational treatment). There is a challenge with conducting E-R analysis only within the treatment arm; the disease risk factors may not only affect efficacy but could also impact drug exposure (baseline-driven ER relationship).[9] The clinical response itself may also affect exposure (response-driven ER relationship). This may result in artifactual E-R relationships that do not necessarily represent the true relationship between drug exposure and the observed response.[10,11] Although this could be evaluated in a multivariate regression analysis to address the interplay among exposure, disease risk factors, and response (efficacy or safety); the multivariate analysis might lead to false conclusions due to incorrect assumptions from the structural model (for example, linear models or models assuming proportional hazards which may be violated in oncology).[12] Moreover, decoupling the exposure-driven and baseline-driven E-R using the multivariate analysis is only possible when data from multiple randomized dose levels is available because when only one dose level is studied, often baseline risk factors and exposure are highly correlated.[9]

Data from the control arm contains valuable information about important baseline and disease characteristics to contextualize the E-R relationship. There are cases where patients in the treatment arm are classified based on exposure (e.g., low to high exposure quartiles) and

clinical outcome is compared to the control arm. For example, if the treatment was shown to provide clinical benefit versus control, confirmation of clinical benefit in the Q1 subgroup versus control would be useful. This comparison needs to account for the difference in the distribution of patient characteristics and disease risk factors and sample size between Q1 and the control arm. Case–control matching is a systematic way to address such issues.[13] Although the use of case-control analysis is popular in observational studies, their use in E-R analyses has rarely been reported.[11,14] The implementation of case–control matching for E-R analysis for the first time was reported by Wang et al.[11] for patients with metastatic gastric cancer (mGC) to reduce the bias introduced by confounding risk factors through balancing them between the Q1 exposure quartile of the treatment arm (i.e., trastuzumab + chemotherapy) and the control arm. After performing the case–control matching, the results indicated that the survival curve of the patients in the Q1 subgroup overlapped with the matching control group, suggesting that patients in the Q1 trastuzumab subgroup did not benefit from the addition of this drug to the chemotherapy backbone. This analysis supported the FDA recommendation on conducting postmarketing evaluation to determine whether a dosing regimen with higher exposure would result in survival improvement.[15]

Herein, we report the application of case–control matching for exposure-efficacy analysis for avelumab, a PD-L1 inhibitor approved for the treatment of several cancers. Data from the JAVELIN-100 was utilized in our analysis. JAVELIN-100 is a phase III study of avelumab in first-line maintenance treatment in patients with urothelial carcinoma. The study demonstrated superiority of avelumab + best supportive care (BSC) versus BSC alone in patients with urothelial carcinoma (UC). An E-R analysis was previously conducted to explore the correlation of avelumab exposure with efficacy and identify important baseline covariates. However, data from the control arm was not used in this analysis.[16] The current work describes a systematic approach for case–control matching for exposure-efficacy analysis by (a) characterizing key disease risk factors for matching, (b) assessment of different case–control matching approaches, and (c) evaluation of the clinical benefit in the Q1 avelumab subgroup versus the matched subgroup from the control arm.

METHODS

Study design

Data used in this work is from the JAVELIN-100, a phase III, randomized, open-label study comparing avelumab + BSC versus BSC alone in adult patients with unresectable locally advanced or metastatic UC after completion of first-line induction platinum-based chemotherapy (JAVELIN-100 ClinicalTrials.gov number, NCT02603432). Patients were randomized in a 1:1 ratio into two arms (total 700 patients, ~350/arm). Arm A (treatment arm): avelumab intravenous (i.v.) infusion at a dose of 10 mg/kg once every 2 weeks (q2w) together with BSC versus arm B (control arm): BSC alone (i.e., patients were cared for as deemed appropriate by the treating physician with no active antitumor therapy).[17]

Exposure-response analysis

The E-R relationship was evaluated using a parametric time-to-event (TTE) model. For the base model development, the distribution of survival versus time was assessed (in the absence of covariates). Then, potential influential covariates, including exposure, laboratory, and clinical covariates, were selected based on graphical exploration. These covariates were later evaluated in the TTE model using the stepwise covariate model (SCM) building procedure approach to form the final model.[16]

Avelumab exposure metrics were evaluated individually. To avoid the impact of disease response on avelumab clearance, early avelumab exposure metrics (e.g., first-dose area under the curve [AUC], cycle 1 day 15 trough concentration [$C_{\text{trough, C1D15}}$], first-dose clearance, and first-dose amount) were estimated using the population PK (PopPK) approach and used as exposure metrics in the E-R (overall survival [OS]) analyses using a univariable approach to identify the most appropriate exposure metric. The selected exposure metric was subsequently evaluated in multivariate analyses leading to the selection of the final model.[16]

To contextualize the E-R relationship with the control arm, subjects in the treatment arm were stratified by avelumab exposure quartiles and the Kaplan–Meier (KM) curves were compared within the different exposure quartiles and with the control arm.

A directed acyclic graph (DAG) was developed using DAGitty, which is a browser-based environment for creating casual diagrams.[18] The DAG depicts causal relationships among dose, exposure, and the outcome of interest (OS). It also represents the relationship between important covariates and both exposure and OS. In causal DAG schematics, variables are represented as nodes and dependence structure between them are shown with arrows.[19] Confounding relationships (i.e., those affecting both exposure metric and the outcome) are shown in red arrows. The impact of factors affecting only PK parameters are shown in black arrows. Assumptions made to simplify the

initial "complete" DAG into intermediate and then final DAGs were outlined.

Case–control matching analysis

Covariate selection

The first step in case–control matching is selecting appropriate covariates. In our analysis, we evaluated the covariates that were determined to be significant contributors to the OS in a parametric TTE model as described above.

Selection of distance measure

Selected covariates were used to compute the "distance," which is a measure of similarity between two individuals from different groups (i.e., one from the avelumab Q1 group and another from the control arm) for the identified set of covariates. Propensity scores, a commonly used metric for distance measure, summarize all the covariates into one scalar, defined as the probability of receiving one of the treatments being compared, given the measured covariates. It allows the construction of matched subgroups based on a similar distribution of covariates without requiring close or exact matches on all individual variables.[20] If two individuals share similar propensity scores, they are considered 'matched'.[10] Logistic regression, the most common method for calculating the propensity score, was utilized in this analysis.[13] The other method to measure the distance in the current analysis was the Mahalanobis metric. The Mahalanobis metric calculates the distance in a multidimensional space, which generally is used for continuous variables and with relatively few covariates.

Selection of matching methods

To enable equal number of subjects between Q1 and control group, case-to-control ratio was set to 1:1 (i.e., exactly 1 match from the control group was required for each subject in the treatment group). In this analysis, both greedy and optimal matching methods were performed using the distance measures mentioned above (a) nearest neighbor matching as the greedy matching method, where the first subject of the control group that meets the criteria for matching is selected, and (b) optimal matching where the overall set of matches are considered to minimize a global distance measure.[13] To choose the best matching method, the percent improvement in the standardized difference between the treatment and control group before and after matching was calculated for each covariate. Case–control matching analysis was conducted in the MatchIt package in R programming language, version 3.6.1 (R Foundation for Statistical Computing).[21]

Investigating the quality of the matching

Graphical and numerical diagnostic approaches were used to investigate the quality of matching from each method. For graphical diagnostics, histograms (continuous variables) and bar graphs (categorical variables) were evaluated before and after matching. Numerical balance diagnostics were conducted via "standardized difference" to check the covariate balance. Standardized difference (d) is the difference in means of each covariate, divided by the standard deviation of the covariate in the treated and untreated groups, as defined in the following equations.

For continuous covariates:

$$d = \frac{\overline{X}_{\text{treatment}} - \overline{X}_{\text{control}}}{\sqrt{\frac{S^2_{\text{treatment}} + S^2_{\text{control}}}{2}}} \times 100,$$

where $\overline{X}$ is the sample mean of the covariate in the treatment and control groups and S^2 denotes the sample variance of the covariate in the treatment and control groups.

For categorical covariates:

$$d = \frac{\hat{P}_{\text{treatment}} - \hat{P}_{\text{control}}}{\sqrt{\frac{\hat{P}_{\text{treatment}}\left(1-\hat{P}_{\text{treatment}}\right) + \hat{P}_{\text{control}}\left(1-\hat{P}_{\text{control}}\right)}{2}}} \times 100,$$

where $\hat{P}$ is the prevalence or mean of the categorical covariate in the treatment and control groups.

Unlike many other statistical tests of hypothesis, the standardized difference is not influenced by sample size, which is useful for investigating the quality of matching given that through the matching process, many subjects (that are not a good match) will be removed from the analysis resulting in different sample sizes before and after matching. Additionally, different matching methods would result in different sample sizes in the matched group.

Other statistical tests were applied before and after matching to investigate the covariate balance and confirm adequate covariate matching (i.e., t-test for the continuous variables and Fisher Exact test for the categorical variables).

Survival analysis

The OS in the lowest avelumab exposure group and their matched subjects from the control group were graphically

evaluated via KM survival plots. Hazard ratio and 95% confidence intervals (CIs) were also reported for different subgroups using the Cox proportional hazards regression model in Survival package within R. The assumption of hazard proportionality was evaluated via assessment of Schoenfeld residuals.

RESULTS

Study design and patient characteristics

A total of 700 patients ($N=350$ from the avelumab + BSC arm and $N=350$ from the BSC alone arm) were enrolled in the JAVELIN-100 study. Randomization was stratified according to the best response to first-line chemotherapy (complete or partial response (CR/PR) vs. stable disease) and to the metastatic site when first-line chemotherapy was initiated (visceral vs. nonvisceral). The study met its primary end point by demonstrating that the addition of avelumab to BSC resulted in a significantly longer OS versus BSC alone. The median OS was 21.4 months (95% CI: 18.9–26.1) in the avelumab + BSC arm, and 14.3 months (95% CI: 12.9–17.9) in the BSC alone arm, stratified hazard ratio for death, 0.69 (95% CI, 0.56–0.86); $p=0.001$.[17]

Identification of covariates of interest

The early exposure metrics including first-dose AUC, $C_{\text{trough, C1D15}}$, first-dose clearance, and first-dose amount were evaluated in the OS base model using a univariable approach to select the most significant exposure metric for stratification. The early exposure metrics (exposure metrics derived from the first dose) were used to minimize the limitations of the assumed dependence between steady-state exposure metrics and response/post-treatment effects; exposure metrics derived from the first dose are more consistent with the true E-R relationship than steady-state exposure metrics. Avelumab $C_{\text{trough, C1D15}}$ was identified as the most appropriate exposure metric.[16] The potential covariates of interest were selected based on the graphical exploration. Covariates of interest that were identified as potentially influential in the previous step were evaluated in the parametric TTE model using the SCM approach; the following covariates were included in the final model: PD-L1 status (yes/no defined by ≥25% baseline expression), baseline metastasis (nonvisceral versus visceral), which was one of the stratification factors of the study, log of baseline lactate dehydrogenase (LDH), and baseline hemoglobin (BHGB). All four identified covariates were included in the case–control matching analysis as significant effect modifiers to OS. In both the treatment and control arms, a trend was observed between lower baseline LDH, higher BHGB, positive PD-L1 expression, and nonvisceral baseline metastasis with higher probability of OS. The second stratification factor (best response to first-line induction chemotherapy [CR/PR vs. SD]) was not significant in the parametric time-to-event model using the SCM approach and was not included in the final model. After screening the subjects for the four covariates listed above, the patients whose PD-L1 status was missing were removed from the analysis which resulted in 328 patients in the treatment (avelumab + BSC) arm and 301 patients in the control (BSC only) arm. The patients in the lowest quartile of $C_{\text{trough, C1D15}}$ ($N=82$) were extracted from the treatment arm. A summary of baseline characteristics of patients in the lowest exposure quartile (Q1, $N=82$), the patients in the remaining higher exposure quartiles (Q2–4, $N=246$), the whole treatment arm ($N=328$), and the whole control arm ($N=301$) are presented in Table 1. Some of the identified baseline covariates (e.g., metastatic site, PD-L1 status, and baseline LDH) demonstrated a certain degree of imbalance between Q1 and the control arm. Moreover, the level of imbalances differs between covariates.

ER assessment for overall survival before matching

The OS for patients in the avelumab + BSC arm stratified by avelumab higher and lower exposure groups was compared to the control arm and is shown in Figure 1. The median OS was 24.1 months in the higher exposure groups (Q2–4) and 14.8 months in the control arm (hazard ratio for death: 0.63, 95% CI: 0.49–0.81, $p=0.0003$), indicating that higher avelumab exposure is associated with longer OS. However, there is a significant overlap of curves for patients in the Q1 avelumab group and the control arm (hazard ratio for death: 0.95, 95% CI: 0.66–1.3, $p=0.772$). The seemingly overlapping survival curves might suggest less clinical benefit in the Q1 avelumab group compared to the control arm. However, as shown previously,[16] several factors may confound the interpretation of exposure-OS results, including imbalance of baseline health status between the Q1 avelumab group and the control arm. An initial "complete" DAG was developed to represent the causal relationships among avelumab dose, exposure, relevant covariates, and OS (Figure 2a).[19] The relationship between dose and OS is assumed to be mediated via $C_{\text{trough, C1D15}}$. Significant covariates on PopPK parameters (e.g., body weight on clearance) are demonstrated. Clinically relevant covariates on OS (i.e., metastatic site, PD-L1 status, baseline hemoglobin, and baseline LDH) were included as contributors to the outcome, OS. Baseline

TABLE 1 Summary of patient's characteristics (significant covariates).

Baseline covariates		First quartile *N* = 82	Combined second to fourth quartiles *N* = 246	Treatment arm *N* = 328	Control arm *N* = 301
Metastatic site, *n*	Nonvisceral	33 (40%)	118 (48%)	151 (46%)	136 (45%)
	Visceral	49 (60%)	128 (52%)	177 (54%)	165 (55%)
PD-L1 status, *n*	Positive	49 (60%)	140 (57%)	189 (58%)	169 (56%)
	Negative	33 (40%)	106 (43%)	139 (42%)	132 (44%)
Baseline hemoglobin, g/L	Mean	118.10	118.23	118.20	117.02
	SD	14.27	13.98	14.03	13.81
	Median	117.00	117.00	117.00	117.00
	Range	85.00–150.00	88.00–159.00	85.00–159.00	10.90–155.00
Baseline lactate dehydrogenase, IU/L	Mean	282.22	258.59	264.50	257.60
	SD	176.04	109.09	129.22	123.08
	Median	224.00	222.00	222.00	225.00
	Range	113.00–1341.00	113.00–968.00	113.00–1341.00	77.00–1680.00

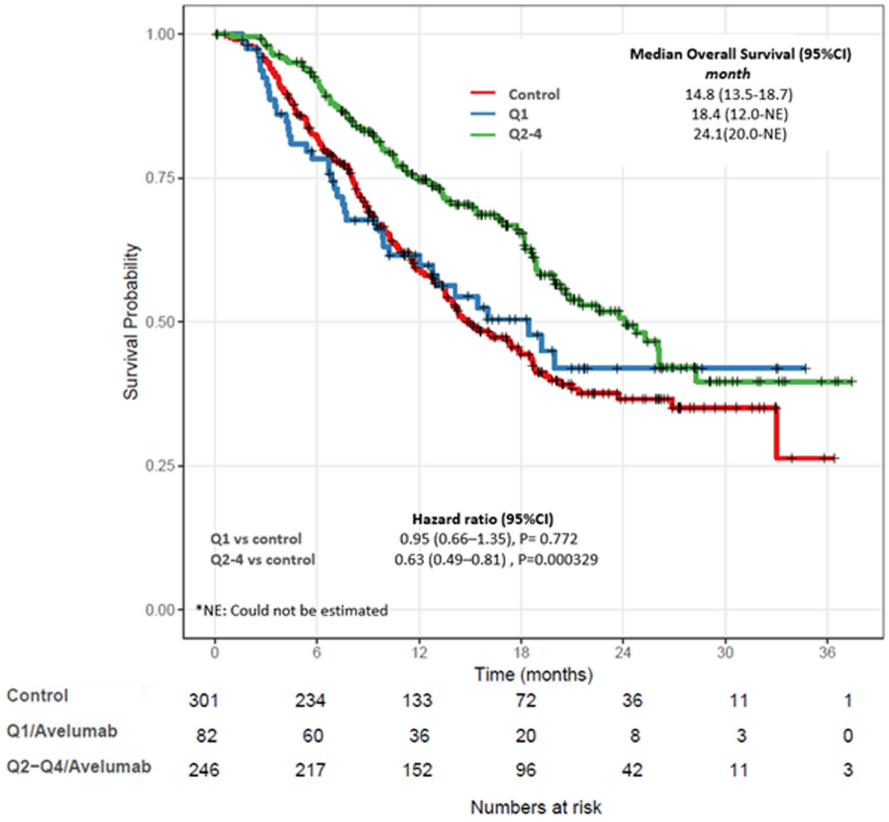

FIGURE 1 Kaplan–Meier overall survival plot. For patients in arm A (avelumab + BSC), stratified by avelumab exposure (cycle 1 day 15 trough concentration) quartiles and arm B (BSC) control. BSC, best-supportive care; CI, confidence interval; NE, not estimated.

disease characteristics is assumed to affect both clearance (i.e., $C_{\text{trough, C1D15}}$) and the four clinically relevant covariates on OS. This may confound the E-R relationship as demonstrated via red arrows. Although baseline tumor burden was not a significant covariate in the E-R-efficacy analysis, its impact on OS is displayed based on the plausible relationship, which may confound the E-R relationship as demonstrated via a red arrow.

Assumptions to further simplify the initial DAG in to an intermediate DAG include: (a) covariates on PKs are accounted for via the use of individual PopPK model predictions of $C_{\text{trough C1D15}}$ (Figure 2b), (b) the impact of baseline tumor burden on OS was considered negligible based on the results of the E-R-efficacy analysis, and (c) case–control matching accounted for the confounding effect of the clinically relevant covariates and minimized

the bias when comparing similar populations between avelumab Q1 and the matched control group (Figure 2c). This enables a direct assessment of the causal effect of $C_{trough,\ C1D15}$ on OS, which is the main objective of this analysis (Figure 2d).

Case–control matching

Propensity score and Mahalanobis methods for distance measure calculation combined with several matching methods were explored using the MatchIt package in R to perform a 1:1 matching, which would result in the extraction of 82 patients from the control arm that best match the 82 patients in the Q1 treatment group. Mahalanobis metric was selected as the distance measure as it performs the best when there are a few covariates in the analysis to match on. This distance measure was then used as the measure of closeness/similarity to conduct the case matching exercise using different methods (greedy and optimal). Nearest neighbor matching (as the greedy method) and optimal full or pair matching (as the optimal method) were performed and the percent improvement in the standardized difference between the Q1 treatment and control group, before and after matching was calculated for each covariate. Although the balance between most covariates improved using the nearest neighbor matching, optimal pair matching had the best results in terms of decreasing the standardized difference after matching. So, the nearest neighbor matching was not further investigated.

As mentioned above, the optimal full matching is a form of optimal pair matching where the participants in either the control or treatment groups will be matched to one or more individuals in the opposite group. To perform a 1:1 matching, a limit for the minimum and the maximum number of control subjects were specified in the code (i.e., set to 1) for the optimal full matching. After applying this limit to the optimal full matching, matching resulted in the same outcome as the optimal pair matching. Using the Mahalanobis optimal matching, 100% improvement in balance was observed in three (out of 4) covariates investigated so it was selected as the best matching method based on the percent improvement in standardized difference values for covariates. *T*-test and Fisher Exact test were performed for continuous and categorical variables, respectively, to compare the *p* value before and after matching. The *p* values along with the percent improvement in standardized difference are presented in Table 2. Better matches resulted in higher *p* value and percent improvements for each covariate. Finally, the survival analysis was performed with the matches that were obtained from the nearest neighbor matching using the Mahalanobis as the distance measure.

E-R analysis for overall survival after matching

The KM curves for the Q1 group and their matched control patients, as well as remaining exposure groups (Q2–4) and the remaining control patients, are presented in Figure 3. A considerable separation in OS curves of Q1 and their matched control is observed. The median OS was 18.4 months in the Q1 avelumab group and 12.1 months in the matched control group (hazard ratio for death in Q1 vs. matched control group is 0.73; which is consistent with that in the remaining avelumab versus remaining control [0.7]), indicating that even the patients in the lowest quartile of avelumab exposure are benefiting from the treatment compared to the control.

DISCUSSION

When performing E-R analysis for oncology drugs, these relationships might be confounded by baseline factors and response to treatment.[6] The potential association between drug exposure, baseline factors, and efficacy may result in an artifactual E-R relationship suggesting that higher exposure leads to better efficacy. However, this does not necessarily represent a causal link between exposure and efficacy and may lead to incorrect dose recommendations.[10,11] This is particularly true for oncology drugs where traditionally only a limited number of patients per dose level are evaluated in dose escalation studies with no dose ranging phase II studies. Disentangling the relationship among drug exposure, baseline factors, disease characteristics, and tumor response might be feasible with data from wide range of doses.

Randomization is the gold standard to balance important baseline factors known to affect the outcome between the treatment and control arms of a clinical trial.[22] Typically, in a controlled study, only subjects from the treatment arm are included in the E-R analysis, as these subjects can be classified based on drug exposure (e.g., Q1 to Q4 of AUC, C_{trough}, or other exposure metrics). In several cases, especially for monoclonal antibodies, the Q1 subgroup might not be perceived to derive sufficient clinical benefit versus the control arm. Classification based on drug exposure within the treatment arm is not randomized and therefore the important baseline factors and disease characteristics between Q1 and the control arm might not be balanced. The sample size differences between Q1 subgroup and the entire control arm might also bias the analysis.

Case–control matching balances the distribution of baseline risk factors between groups. This approach aims to identify a subgroup of subjects within the control arm with similar/balanced baseline factors to subjects in the Q1 group. Consequently, the efficacy outcome in this subgroup from the control arm is compared to that in individuals from the Q1 group.

This type of analysis is not merely an academic exercise; it has been used by the FDA to support regulatory decisions. The analysis by Wang et al. was implemented for patients with mGC to compare the survival of patients treated with trastuzumab and fluoropyrimidine and cisplatin (FC) versus patients who received FC only.[11] The purpose of this analysis was to reduce the bias introduced

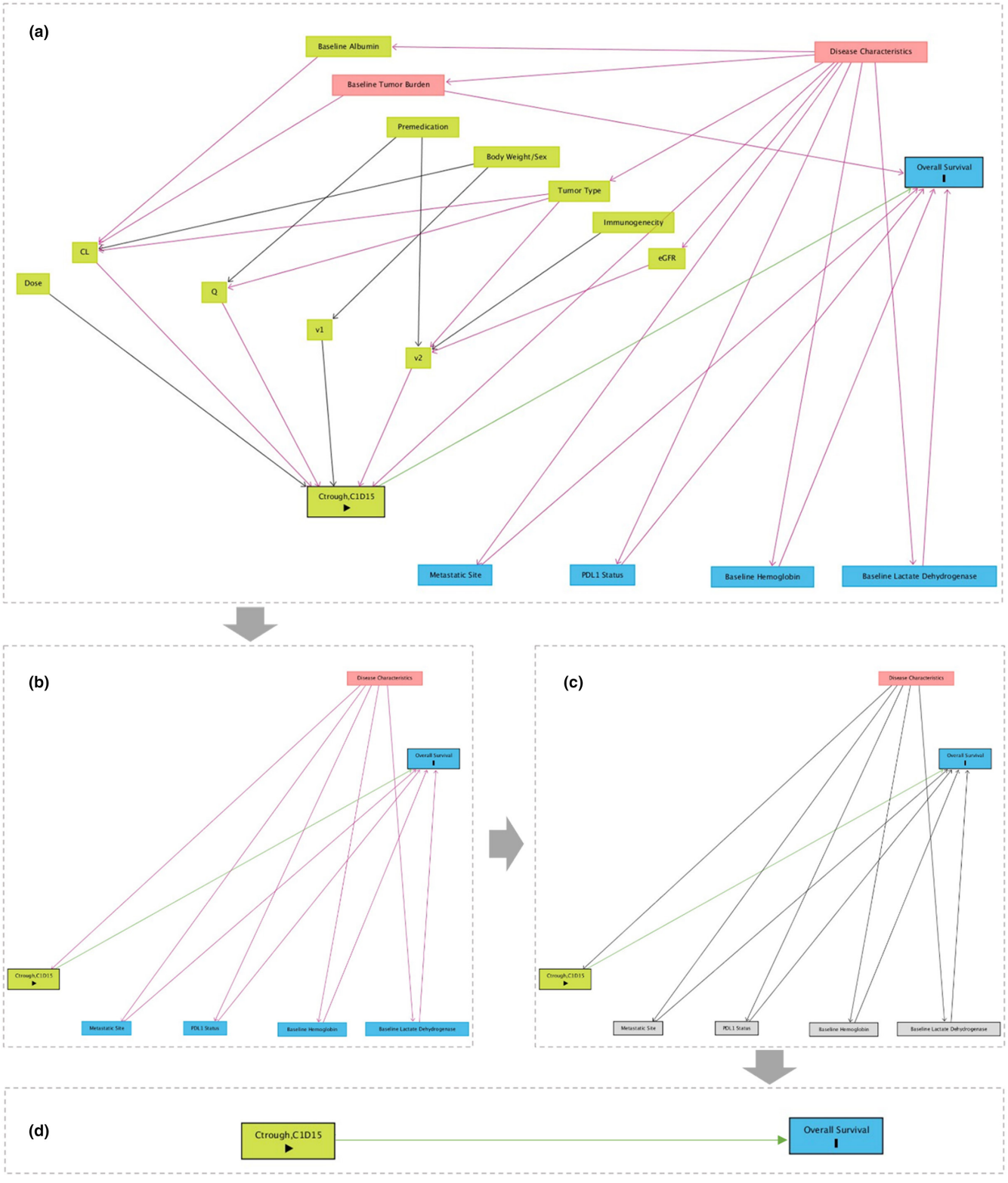

FIGURE 2 DAGs. (a) The initial DAG representing the causal relationship between avelumab dose, exposure, relevant covariates, and OS. Baseline albumin, baseline tumor burden, pre-medication, body weight, sex, tumor type, immunogenicity, and eGFR were identified as significant covariates on PopPK parameters (CL, Q, V1, and V2). The effect of dose is mediated through $C_{trough,\ C1D15}$. Metastatic site, PD-L1 status, baseline hemoglobin, and baseline LDH were included as the significant contributions to the OS based on the parametric time-to-event model. The confounding paths to the E-R relationship (i.e., variables affecting both the exposure and the outcome – mainly baseline disease characteristics) are shown in red arrows. Although baseline tumor burden was not a significant covariate in the E-R-efficacy analysis, its impact on OS is displayed based on the plausible relationship. (b) Intermediate DAG demonstrating a simplified representation of causal relationships by assuming that the effect of PK parameters is accounted for through the use of individual PopPK predictions in the E-R analyses. The impact of baseline tumor burden on OS was considered negligible based on the results of the E-R-efficacy analysis. (c) Another intermediate DAG is represented by further assuming that the confounding effect of the four covariates affecting OS is addressed by case–control matching. (d) This final DAG represents the direct assessment of causal effect of exposure metric ($C_{trough,\ C1D15}$) on response (OS). CL, clearance; C_{trough}, trough plasma concentration; DAG, directed acyclic graph; eGFR, estimated glomerular filtration rate; E-R, exposure-response; OS, overall survival; PK, pharmacokinetic; PopPK, population pharmacokinetic; Q, intercompartmental clearance.

TABLE 2 Balance of covariates before and after matching using the Mahalanobis optimal matching.

	p value		Standardized difference		Improvement in standardized difference (%)
	Before matching	After matching	Before matching	After matching	
Metastatic site (nonvisceral vs. visceral)	0.45	1.00	10.00	0.00	100
PD-L1 status	0.61	1.00	7.30	0.00	100
Baseline hemoglobin	0.54	1.00	7.70	0.00	100
Baseline lactate dehydrogenase	0.24	0.84	16.20	3.20	80.25

by confounding risk factors through balancing them between the Q1 group and the control group. After performing the case–control matching, the exposure survival analysis was conducted to compare the survival of the Q1 group to the matching control group. In that example, the results indicated that the survival curve of the patients in the lowest-quartile trough concentration (C_{min}) of trastuzumab in cycle 1 mainly overlapped with the matching control group, suggesting that these patients did not benefit from the addition of trastuzumab to FC. This finding supported the FDA recommendation on conducting post-marketing trials to investigate whether higher dose increases trastuzumab C_{trough} levels and increases OS.[4]

In another study, case–control matching analysis was reported to compare the progression-free survival and OS of trastuzumab emtansine (T-DM1) versus capecitabine plus lapatinib (control) in previously treated human epidermal growth factor receptor 2-positive advanced breast cancer using data from the EMILIA study.[14] Various exposure metrics including model-predicted C_{min} and AUC as well as observed AUC and maximum plasma concentration were tested in this analysis. The results indicated that the E-R relationships were most likely confounded by both the patients' baseline risk factors and the choice of exposure metrics that was used in the analysis.[4]

In another work, the case–control matching was used to investigate the effect of time-varying clearance of nivolumab on disease dynamics and further E-R analysis.[23] To perform case–control analysis, patients in the treatment arm were divided into four quartiles by drug exposure which later were matched with the subjects extracted from the control arm. Three different exposure metrics were used (average concentration at steady-state, average concentration at cycle 1, and trough concentration at cycle 1), each of which resulted in different E-R conclusions. Using the case–control matching approach and balancing the distribution of baseline risk factors across the exposure quartiles, it was concluded that early exposure metrics (after the first dose) are less confounded by differences in prognostic factors and the E-R relationship appeared much flatter.[4]

In this work, causal DAGs were used to evaluate the relationships between dose, exposure metrics, potentially influential covariates, and the efficacy outcome. The DAG represents an explicit representation of the causal assumptions governing the interplay between exposure and outcome. It was also used to represent the simplifying assumptions from a complete DAG to a final one enabling causal assessment of the relationship between avelumab exposure and OS. These DAGs were helpful in elucidating potential biases and approaches to simplify or address these biases.[19]

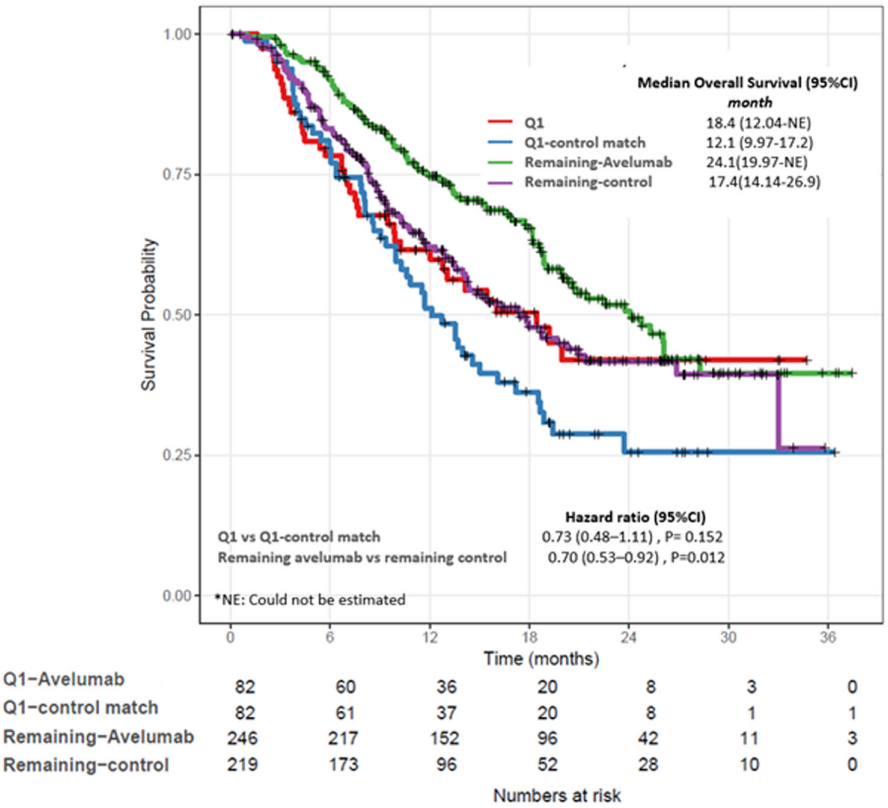

FIGURE 3 Kaplan–Meier curves for the lower exposure quartile (Q1) and higher exposure quartiles combined (Q2–4) and the matched control and remaining control groups. CI, confidence interval; NE, not estimated.

The selection of the case–control matching method to optimize the matching process as well as the criteria for which covariates to include in matching are still controversial topics and were not well-described in the published works mentioned above. In our analysis, we included the covariates that were significant contributors to the efficacy outcome (OS) in a parametric TTE analysis. To define the distance measure, we investigated both the propensity score and Mahalanobis that are most frequently used in the literature. The main difference between these two distance measures is the approach of weighing covariates. In propensity score matching, the covariates are not equally important, meaning covariates that have a stronger relationship with the treatment group, are weighted more heavily than covariates that have a weaker relationship with the treatment group[21]; whereas Mahalanobis distance matching equally balances all covariates.

In general, Mahalanobis performs better with few numbers of covariates to match on, especially with normally distributed covariates, whereas the propensity score method relies on the estimated propensity score for matching, rather than the distribution of the covariates.[13] However, determining the best distance measure for each dataset relies on the evaluation of the match quality after the matching process is performed.

After using each distance measure and matching methods, the percent improvement in standardized difference was calculated for each covariate. The selected optimal matching method worked well in reducing the standardized difference of covariates for all the covariates, resulting in 100% improvement for three covariates. This is probably because the size of the control arm was large enough to find the matches for the Q1 of exposure. In addition, the calculated *p* value demonstrated an increase after matching which is indicator of less significant difference between the two groups (i.e., better balance after matching). Similar to what was reported by Wang et al.,[11] the *p* value of "1" was achieved for most of the covariates in our analysis after the matching except for one of them (baseline LDH). This difference is likely due to matching to both categorical and continuous variables in our analysis versus solely on categorical variables by Wang et al. To check this hypothesis, our analysis was repeated again, this time including only categorical variables and the result showed *p* value of "1" for all the covariates similar to Wang's work. After adjusting for the four baseline risk factors by case-matching analysis, the OS curve for patients in the Q1 subgroup was separated from the matched control group indicating that the patients (even being in the lowest avelumab exposure group) are drawing benefit from treatment compared to the matched control group. These types of E-R analyses may guide the use of biologics in different patient subgroups. This information can be helpful for the patients, prescribers, and payors as they navigate different treatment options.

One of the challenges in implementation of case–control matching methods for E-R analyses is dealing with the small sample sizes. For example, the total number for the control arm in this example is 320 patients, which may not be large enough for adequate matching. However, the ratio of the control group sample size compared to the treatment group sample size (e.g., ratio = 4 [320/82] in the avelumab example) could be more influential in matching than the total sample size.[24] In general, it could be challenging to tease out the effect of sample size and control-to-treatment group ratio because the initial difference in the distribution of covariates between treatment and control groups heavily influences whether or not adequate matches can be found. Real-word evidence/real-word data can be potentially used to build a larger and richer analysis dataset to capture the data that are not collected in the context of conventional randomized controlled trials and improve the case–control matching performance.

In conclusion, the case–control matching method was implemented in our E-R analysis to build a dataset including the patients from both the treatment and control arms with similar baseline characteristics. Given that this analysis is based on one dose level, case–control analysis may be still confounded leading to apparent E-R and the analysis is not suitable to draw inferences regarding optimal dose of avelumab but rather the purpose of the analysis here is to show that avelumab is providing benefit in low quartile group compared to matched patients in the BSC arm. The use of case-matching can be an important tool to reduce the bias in oncology E-R analyses that, despite its wide utility, are susceptible to being confounded.

AUTHOR CONTRIBUTIONS

P.S. and M.E. wrote the manuscript. M.E., P.S., A.A., J.H., D.D.W., C.B., and S.R. designed the research. M.E. and P.S. performed the research. P.S. and J.H. analyzed the data.

ACKNOWLEDGMENTS

The authors thank Dr. James A. Rogers from Metrum Research Group for his insights and discussions on developing the DAG presented in this manuscript.

FUNDING INFORMATION

No funding was received for this work.

CONFLICT OF INTEREST STATEMENT

P.S., M.E., J.H., S.R., D.W., C.B., are employees of Pfizer and receive stock and stock options as part of their employment. A.A. was an employee of Pfizer and received stock and stock options as part of his employment at the time of conducting this work.

ORCID

Pooneh Soltantabar https://orcid.org/0000-0002-1396-113X
Mohamed Elmeliegy https://orcid.org/0000-0002-9825-5890

REFERENCES

1. FDA's project optimus. Accessed September 1, 2023. https://www.fda.gov/about-fda/oncology-center-excellence/project-optimus
2. Moran-McCabe K, Burris S. Eviction and the necessary conditions for health. *N Engl J Med*. 2021;385:1443-1445. doi:10.1056/NEJMp2031947
3. Blumenthal G. Optimizing dosing in oncology drug development. 1–14 2021.
4. Soltantabar P, Lon HK, Parivar K, Wang DD, Elmeliegy M. Optimizing benefit/risk in oncology: review of post-marketing dose optimization and reflections on the road ahead. *Crit Rev Oncol Hematol*. 2023;182:103913. doi:10.1016/j.critrevonc.2023.103913
5. Mittapalli RK, Guo C, Drescher SK, Yin D. Oncology dose optimization paradigms: knowledge gained and extrapolated from approved oncology therapeutics. *Cancer Chemother Pharmacol*. 2022;90:207-216. doi:10.1007/s00280-022-04444-0
6. Kawakatsu S, Bruno R, Kågedal M, et al. Confounding factors in exposure–response analyses and mitigation strategies for monoclonal antibodies in oncology. *Br J Clin Pharmacol*. 2021;87:2493-2501. doi:10.1111/bcp.14662
7. Wang Y, Booth B, Rahman A, Kim G, Huang SM, Zineh I. Toward greater insights on pharmacokinetics and exposure–response relationships for therapeutic biologics in oncology drug development. *Clin Pharmacol Ther*. 2017;101:582-584. doi:10.1002/cpt.628
8. Good E, Practice C. Guidance for industry. *Fed Regist*. 2004;505:79.
9. Dai HI, Vugmeyster Y, Mangal N. Characterizing exposure–response relationship for therapeutic monoclonal antibodies in Immuno-oncology and beyond: challenges, perspectives, and prospects. *Clin Pharmacol Ther*. 2020;108:1156-1170. doi:10.1002/cpt.1953
10. Evaluation M, Olmos A. Propensity scores: a practical introduction using R. *J Multidiscip Eval*. 2015;11:68-88. doi:10.56645/jmde.v11i25.431
11. Yang J, Zhao H, Garnett C, et al. The combination of exposure-response and case-control analyses in regulatory decision making. *J Clin Pharmacol*. 2013;53:160-166. doi:10.1177/0091270012445206
12. Rulli E, Ghilotti F, Biagioli E, et al. Assessment of proportional hazard assumption in aggregate data: a systematic review on statistical methodology in clinical trials using time-to-event endpoint. *Br J Cancer*. 2018;119:1456-1463. doi:10.1038/s41416-018-0302-8
13. Stuart EA. Matching methods for causal inference: a review and a look forward. *Stat Sci*. 2010;25:1-21. doi:10.1214/09-STS313
14. Li C, Wang B, Chen SC, et al. Exposure–response analyses of trastuzumab emtansine in patients with HER2-positive advanced breast cancer previously treated with trastuzumab and

a taxane. *Cancer Chemother Pharmacol*. 2017;80:1079-1090. doi:10.1007/s00280-017-3440-4
15. Shah MA, Xu RH, Bang YJ, et al. HELOISE: phase IIIb randomized multicenter study comparing standard-of-care and higher-dose trastuzumab regimens combined with chemotherapy as first-line therapy in patients with human epidermal growth factor receptor 2–positive metastatic gastric or gast. *J Clin Oncol*. 2017;35:2558-2567. doi:10.1200/JCO.2016.71.6852
16. Hibma J, Li J, Bello C, et al. Exposure-response efficacy analysis of avelumab in patients with locally advanced or metastatic urothelial carcinoma (UC) in JAVELIN Bladder 100. *Cancer Res*. 2021;29:2021. doi:10.1158/1538-7445.AM2021-1360
17. Powles T, Park SH, Voog E, et al. Avelumab maintenance therapy for advanced or metastatic urothelial carcinoma. *N Engl J Med*. 2020;383:1218-1230. doi:10.1056/NEJMoa2002788
18. DAGitty. Accessed June 1, 2023. http://www.dagitty.net/
19. Rogers JA, Maas H, Pitarch AP. An introduction to causal inference for pharmacometricians. *CPT Pharmacometrics Syst Pharmacol*. 2023;12:27-40. doi:10.1002/psp4.12894
20. Copas JB, Malley PF. A robust. 4267–4278 2008. doi:10.1002/sim
21. Sam X, Rosenbaum PR. Comparison of multivariate matching methods: structures, distances, and Al Source. *J Comput Graph Stat*. 2016;2:405-420. doi:10.1080/10618600.1993.10474623
22. Bonell CP, Hargreaves J, Cousens S, et al. Alternatives to randomisation in the evaluation of public health interventions: design challenges and solutions. *J Epidemiol Community Health*. 2011;65:582-587. doi:10.1136/jech.2008.082602
23. Liu C, Yu J, Li H, et al. Association of time-varying clearance of nivolumab with disease dynamics and its implications on exposure response analysis. *Clin Pharmacol Ther*. 2017;101:657-666. doi:10.1002/cpt.656
24. Jacovidis J. Evaluating the performance of propensity score matching methods: a simulation study. Diss. 2014–2019 2017. Accessed June 1, 2023. https://commons.lib.jmu.edu/diss201019/149

How to cite this article: Soltantabar P, Alhadab A, Hibma J, et al. Case–control matching-guided exposure-efficacy relationship for avelumab in patients with urothelial carcinoma. *CPT Pharmacometrics Syst Pharmacol*. 2023;12:2001-2012. doi:10.1002/psp4.13049